TIPS & TRICKS FOR GETTING MOST VALUE OUT OF NUMERICAL DATA

Based on a manuscript draft written with Sabrina Rüschenbaum

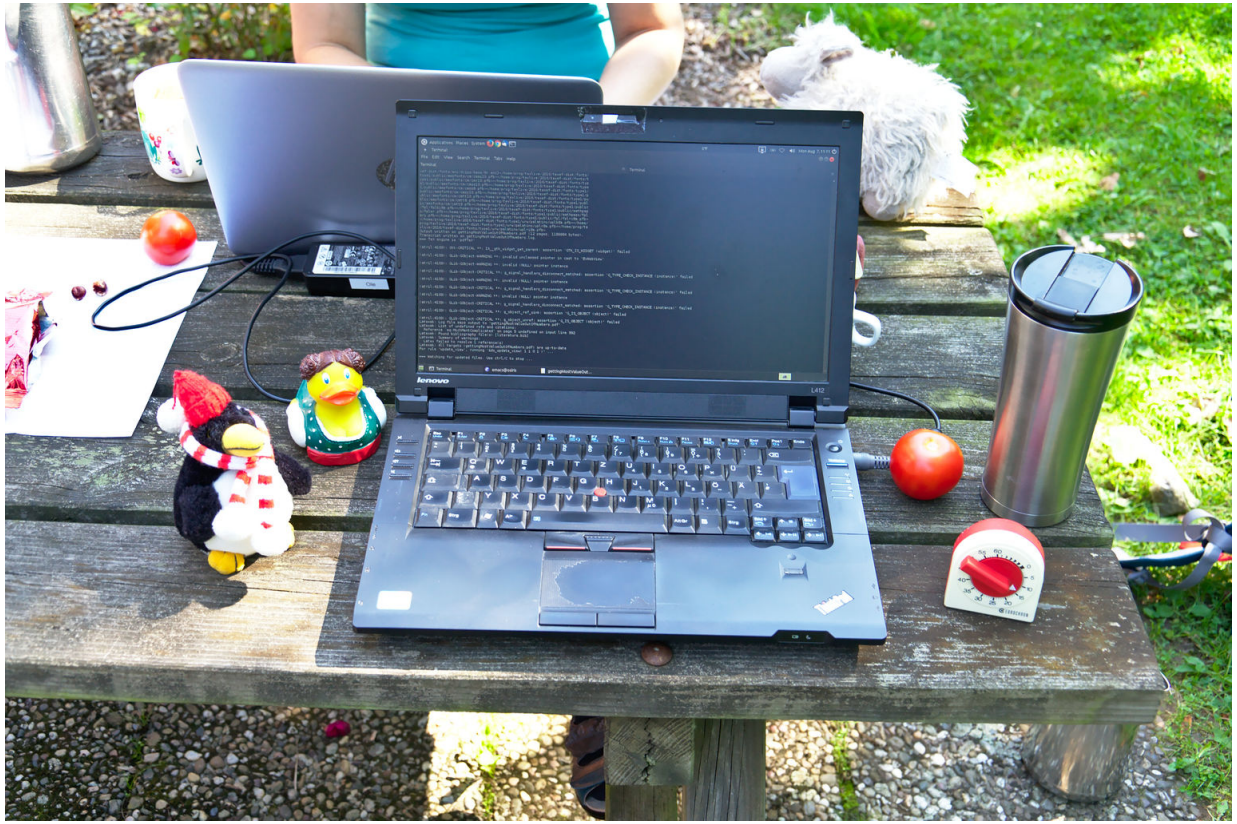
Lars Ole Schwen

CdE Biomodels Retreat 2017, Caldes d'Estrac/Barcelona, 2017-09-13

TIPS & TRICKS FOR GETTING MOST VALUE OUT OF NUMERICAL DATA

1. Introduction

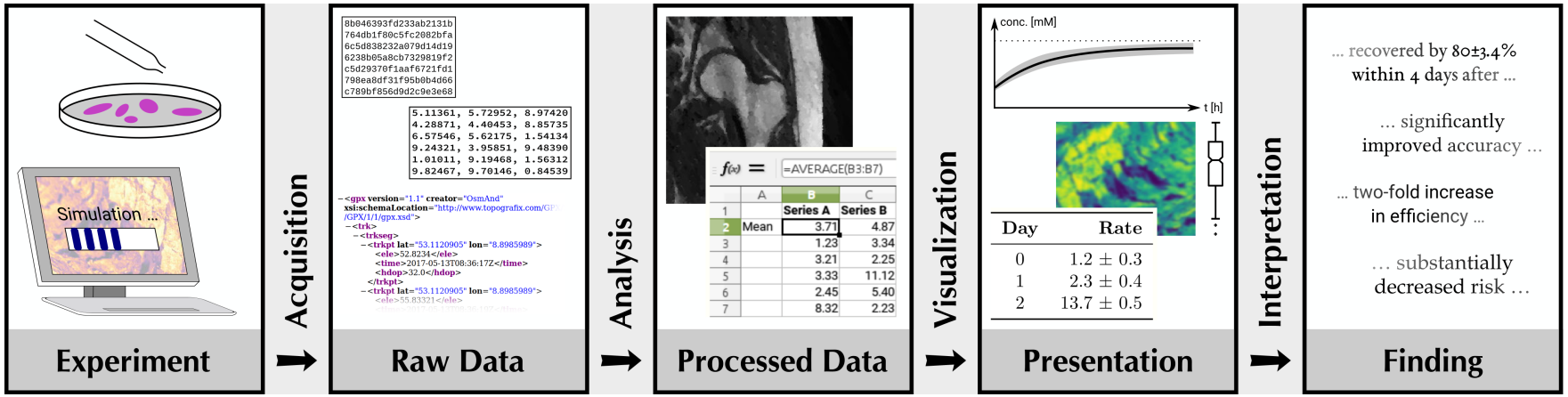
Getting Things Finished Workshop @ CdE SommerAkademie 2017



- mostly written this summer

1. Introduction

Typical Scientific Workflow



1. Introduction

Basics

- numerical data frequently basis for findings

1. Introduction

Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible

1. Introduction

Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible
- mistakes are annoying, but hard to avoid

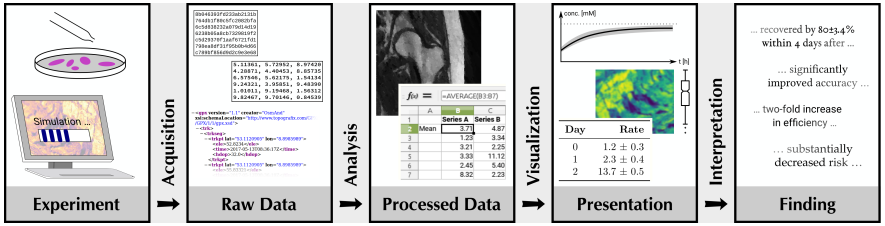
1. Introduction

Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible
- mistakes are annoying, but hard to avoid
- try to make life easy for yourself and others

2. Keep Data Correct

Contents



2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

- two out of every seven values zero
- count of colony-forming units 15, 46, 28, 46, 52, and 46
- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46

- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$
  - measurement every hour or computed value?

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

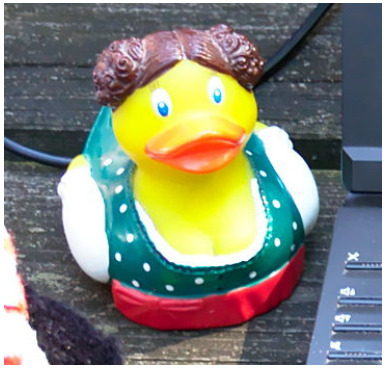
- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$
  - measurement every hour or computed value?
- discuss data with other person

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\mathrm{s}$ and $t_2 = 20\,021.392\,\mathrm{s}$
  - measurement every hour or computed value?
- discuss data with other person
  (or explain it to your rubber duck)



2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Practice guesstimation

- simplified assumptions and estimates
- back-of-the-envelope calculations
- examples below …
- Further reading: https://what-if.xkcd.com

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

- 5 × 2 kW heating for three weeks

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

- 5 × 2 kW heating for three weeks
- 10 kW means 200 kWh per day (20 hours),

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

- 5 × 2 kW heating for three weeks
- 10 kW means 200 kWh per day (20 hours),
- i.e., 6000 kWh in three weeks (3 times 10 days)

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

- 5 × 2 kW heating for three weeks
- 10 kW means 200 kWh per day (20 hours),
- i.e., 6000 kWh in three weeks (3 times 10 days)
- i.e., 1500 € (25 ct/kWh)

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Prospective example: temperature-dependency of CO_2 uptake of plants in a greenhouse

- 5 × 2 kW heating for three weeks
- 10 kW means 200 kWh per day (20 hours),
- i.e., 6000 kWh in three weeks (3 times 10 days)
- i.e., 1500 € (25 ct/kWh)
- budget problem or negligible?

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Retrospective example:

- observed growth of new-born Humboldt penguins for one year.

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Retrospective example:

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Retrospective example:

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Retrospective example:

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

✘ 50 weeks × 5 cm/week = 2.5 m

2. Keep Data Correct

Tip 1: Plausibility-Check Your Data

Retrospective example:

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

✘ 50 weeks × 5 cm/week = 2.5 m

Other example:

- Flight from Toronto to Sydney in 2.5 hours

2. Keep Data Correct

Tip 2: Track Your Sources

2. Keep Data Correct

Tip 2: Track Your Sources

Pitfalls

- Undocumented way how or in which context a number was measured
- Will you find “the snail paper by Smith” in two weeks?
- Will you find $v_{\text{mean}} = 1.27$ millimeters per second if it was originally written as $V_{\text{average}} = 3$ inches per minute?

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again."
  (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again."
  (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

- scripts (R, python, …) better than spreadsheets

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again."
  (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

- scripts (R, python, …) better than spreadsheets
- also avoids data overwriting and formula range problems

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again."
  (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

- scripts (R, python, …) better than spreadsheets
- also avoids data overwriting and formula range problems
- document what you did

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again."
  (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

- scripts (R, python, …) better than spreadsheets
- also avoids data overwriting and formula range problems
- document what you did
- version-control evaluation (and manuscripts)

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: “84/96 cases OK” always produced zero

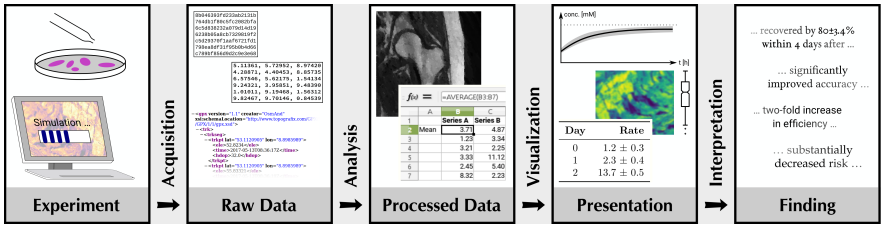
2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: “84/96 cases OK” always produced zero
- Non-finite values (NaN)

3. Correctly Interpret Your Data

Contents



3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

- What is “temperature in the 30s”?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

- What is “temperature in the 30s”?
  - summer day at 30°C in Paris
  - expect snow at 30°F in Seattle
  - superconductor experiment at 30 K

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

- What is “temperature in the 30s”?
  - summer day at 30°C in Paris
  - expect snow at 30°F in Seattle
  - superconductor experiment at 30 K
- “A [mathematical] model without units is not a model.” (Matthias König)

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed due to using pound seconds instead of Newton seconds as the unit of impulse

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed due to using pound seconds instead of Newton seconds as the unit of impulse
- Hochrheinbrücke crossing the river Rhine originally planned with sign error when compensating for “above sea level” references CH vs. DE

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed due to using pound seconds instead of Newton seconds as the unit of impulse
- Hochrheinbrücke crossing the river Rhine originally planned with sign error when compensating for “above sea level” references CH vs. DE

Caldes d’Estrac, train station

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read
  - erythrocyte diameter is 7 µm

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive
- 8×10^{-8} m² fuel consumption of a car correct SI unit and interesting to interpret, but hard to judge

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m harder to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive
- 8×10^{-8} m² fuel consumption of a car correct SI unit and interesting to interpret, but hard to judge
  - 8 liters per 100 kilometers (or fuel efficiency of 29.4 miles per gallon) better

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know thy SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know thy SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know thy SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?
  - one billion = Eine Billion?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know thy SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?
  - one billion = Eine Billion?
- calories vs. kcal

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- “two meters plus half a second” does not make sense

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- “two meters plus half a second” does not make sense
- differentiation corresponds to division by unit

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- “two meters plus half a second” does not make sense
- differentiation corresponds to division by unit
  - e.g., velocity in units meters per second is the temporal gradient (one over seconds) times position (meters)

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- “two meters plus half a second” does not make sense
- differentiation corresponds to division by unit
  - e.g., velocity in units meters per second is the temporal gradient (one over seconds) times position (meters)
- integration corresponds to multiplication by unit

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = \dot{c}(t) = -\frac{V_{\mathrm{max}} \cdot c(t)}{K_m + c(t)} \tag{1}$$

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = \dot{c}(t) = -\frac{V_{\mathrm{max}} \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = \dot{c}(t) = -\frac{V_{\mathrm{max}} \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume
- addition in the denominator implies that K_m is a constant also in units of concentration.

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = \dot{c}(t) = -\frac{V_{\mathrm{max}} \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume
- addition in the denominator implies that K_m is a constant also in units of concentration.
- factor c (concentration) in numerator implies that V_{max} has units of concentration per time, i.e., a rate of concentration change.

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Separate Unit Conversion and Formula Logic, or else …

$$\dot{c}(t) = -\frac{\frac{1\,\mathrm{min}\cdot 1\,\mathrm{mol}\cdot \tilde{V}_{\mathrm{max}}}{60\,\mathrm{s}\cdot 180.164\,\mathrm{g}} \cdot c(t)}{\frac{1\,\mathrm{min}\cdot 1\,\mathrm{mol}\cdot \tilde{K}_m}{60\,\mathrm{s}\cdot 180.164\,\mathrm{g}} + c(t)} \tag{2}$$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \rightarrow \pm\infty$ or limits of useful ranges

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$
- differences $\to 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$
- differences $\to 0$
- denominators $\to 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$
- differences $\to 0$
- denominators $\to 0$
- What is the monotonicity in these cases?

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V_{\text{max}} \cdot c(t)}{K_m + c(t)} \tag{3}$$

V_{max} and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V_{\text{max}} \cdot c(t)}{K_m + c(t)} \tag{3}$$

V_{max} and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V_{\max} \cdot c(t)}{K_m + c(t)} \tag{3}$$

$V_{\max}$ and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain $V_{\max}$ (i.e., maximal rate)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V_{\max} \cdot c(t)}{K_m + c(t)} \tag{3}$$

$V_{\max}$ and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain $V_{\max}$ (i.e., maximal rate)
- larger $V_{\max}$: larger decay if all other values stay the same

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V_{\max} \cdot c(t)}{K_m + c(t)} \tag{3}$$

$V_{\max}$ and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain $V_{\max}$ (i.e., maximal rate)
- larger $V_{\max}$: larger decay if all other values stay the same
- K_m more difficult to interpret: $K_m = c$ implies half maximal rate

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Suppose you are planning drug-drug interaction test where each drug purchase requires a lot of bureaucracy.
Effort

- *n* times purchase bureaucracy (*n* number of drugs)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Suppose you are planning drug-drug interaction test where each drug purchase requires a lot of bureaucracy.
Effort

- n times purchase bureaucracy (n number of drugs)
- $n(n-1)/2 = (n^2 - n)/2$ pairs of drugs (experimental effort)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Suppose you are planning drug-drug interaction test where each drug purchase requires a lot of bureaucracy.
Effort

- n times purchase bureaucracy (n number of drugs)
- $n(n-1)/2 = (n^2 - n)/2$ pairs of drugs (experimental effort)
- eventually, n^2 effort will dominate

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

✘ soil quality index (high presence $\not\Rightarrow$ high quality)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

✘ soil quality index (high presence $\not\Rightarrow$ high quality)

✔ soil pollution index (high presence $\Rightarrow$ high pollution)

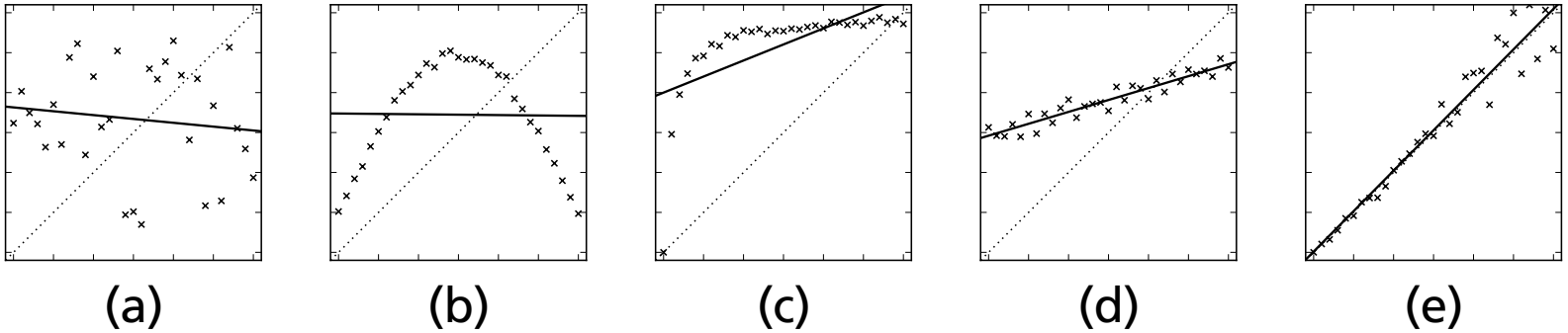
3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

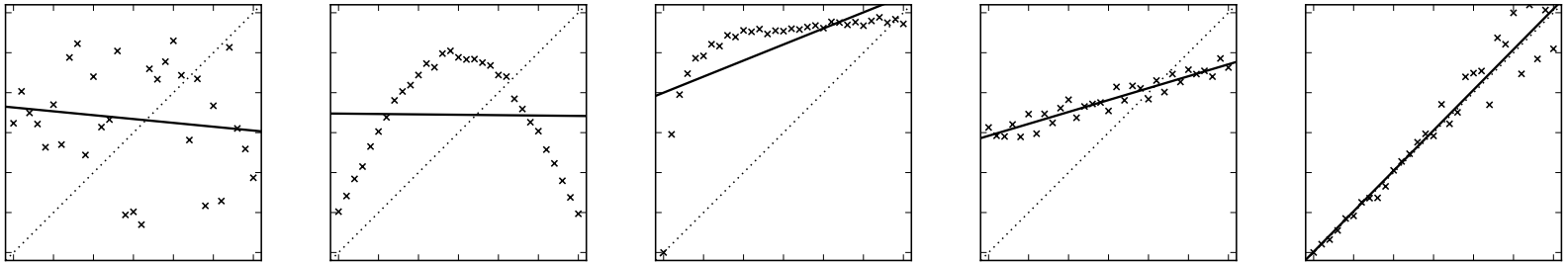
Correlation Coefficients



3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Correlation Coefficients



(a) (b) (c) (d) (e)

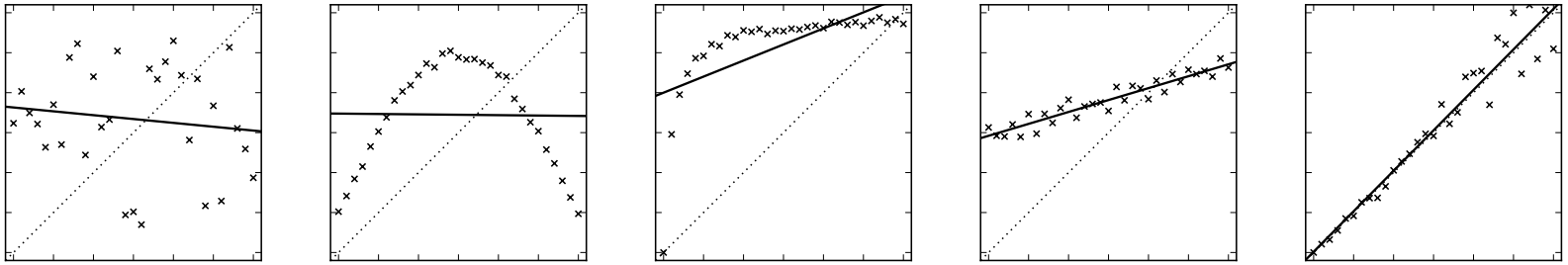
Data Series	Pearson CC
(a) noise	–0.128
(b) non-monotonic	–0.013
(c) monotonic, non-linear	0.639
(d) linear	0.948
(e) identical	0.972

- correlation is not causation, e.g., decreasing number of pirates and global warming [B. Henderson, The Gosple of the Flying Spaghetti Monster])

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Correlation Coefficients



(a) (b) (c) (d) (e)

Data Series	Spearman CC	Pearson CC	Concordance CC
(a) noise	–0.108	–0.128	–0.120
(b) non-monotonic	–0.019	–0.013	–0.012
(c) monotonic, non-linear	0.955	0.639	0.280
(d) linear	0.951	0.948	0.459
(e) identical	0.971	0.972	0.971

- correlation is not causation, e.g., decreasing number of pirates and global warming [B. Henderson, The Gosple of the Flying Spaghetti Monster])

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

- ✔ a high probability that an effect (e.g., difference) is not just due to randomness
- ✘ not proof of the above
- ✘ not a measure how big the effect is

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

- ✔ a high probability that an effect (e.g., difference) is not just due to randomness
- ✘ not proof of the above
- ✘ not a measure how big the effect is
- ✘ not a measure of importance/meaning/relevance

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

- ✔ a high probability that an effect (e.g., difference) is not just due to randomness
- ✘ not proof of the above
- ✘ not a measure how big the effect is
- ✘ not a measure of importance/meaning/relevance
- ✘ not a check whether the test is applicable in the first place

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

- ✔ a high probability that an effect (e.g., difference) is not just due to randomness
- ✘ not proof of the above
- ✘ not a measure how big the effect is
- ✘ not a measure of importance/meaning/relevance
- ✘ not a check whether the test is applicable in the first place
- ✘ lack of significance $\not\Rightarrow$ no meaningful difference

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	Group A			Group B			approx. *p*-value
	group	*n*	Med	group	*n*	Med	
Birth Month							
b-CS	June			Sep.			1.32×10^{-4}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	Group A			Group B			approx. *p*-value
	group	*n*	Med	group	*n*	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	Group A			Group B			approx. *p*-value
	group	*n*	Med	group	*n*	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	Group A			Group B			approx. *p*-value
	group	*n*	Med	group	*n*	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}
Death Day							
b-SH	day 9	168	19.5	day 27	168	6.0	3.02×10^{-5}
b-H	day 9	248	56.5	day 27	250	23.0	3.95×10^{-5}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	Group A			Group B			approx. *p*-value
	group	*n*	Med	group	*n*	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}
Death Day							
b-SH	day 9	168	19.5	day 27	168	6.0	3.02×10^{-5}
b-H	day 9	248	56.5	day 27	250	23.0	3.95×10^{-5}
Birth State							
b-G	CA	2182	121.5	PA	660	33.0	6.74×10^{-27}
b-G	CA	2182	121.5	MA	1410	52.0	1.54×10^{-25}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.

⇒ Also domain non-experts can find seemingly “highly significant” results

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.

⇒ Also domain non-experts can find seemingly "highly significant" results

- This is misuse of reasonable statistical methods for "data dredging" or "*p*-value hacking"

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.

⇒ Also domain non-experts can find seemingly "highly significant" results

- This is misuse of reasonable statistical methods for "data dredging" or "*p*-value hacking"
- At least, Bonferroni correction should be applied (rendering most results above non-significant)

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

- 68.432 853 percent of all statistics pretend to be more accurate than they actually are

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)
- errors propagate through calculations

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)
- errors propagate through calculations
- results should not be based on noise

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Example of Uncertainty Quantification: (low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Example of Uncertainty Quantification: (low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05$ %, assuming no correlation and normal distribution
- distance error dominates

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Example of Uncertainty Quantification: (low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05$ %, assuming no correlation and normal distribution
- distance error dominates

Monte-Carlo Simulations

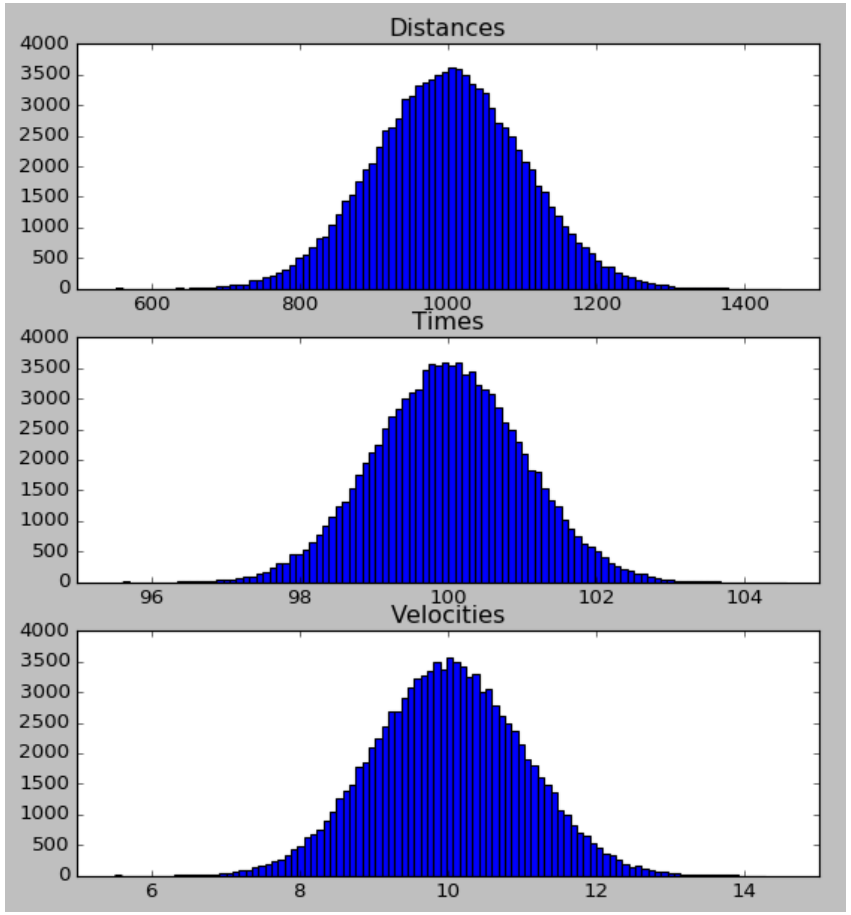
- repeat data evaluation many times with input data sampled from distribution estimating uncertainty

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Example of Uncertainty Quantification: (low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05$ %, assuming no correlation and normal distribution
- distance error dominates



Monte-Carlo Simulations

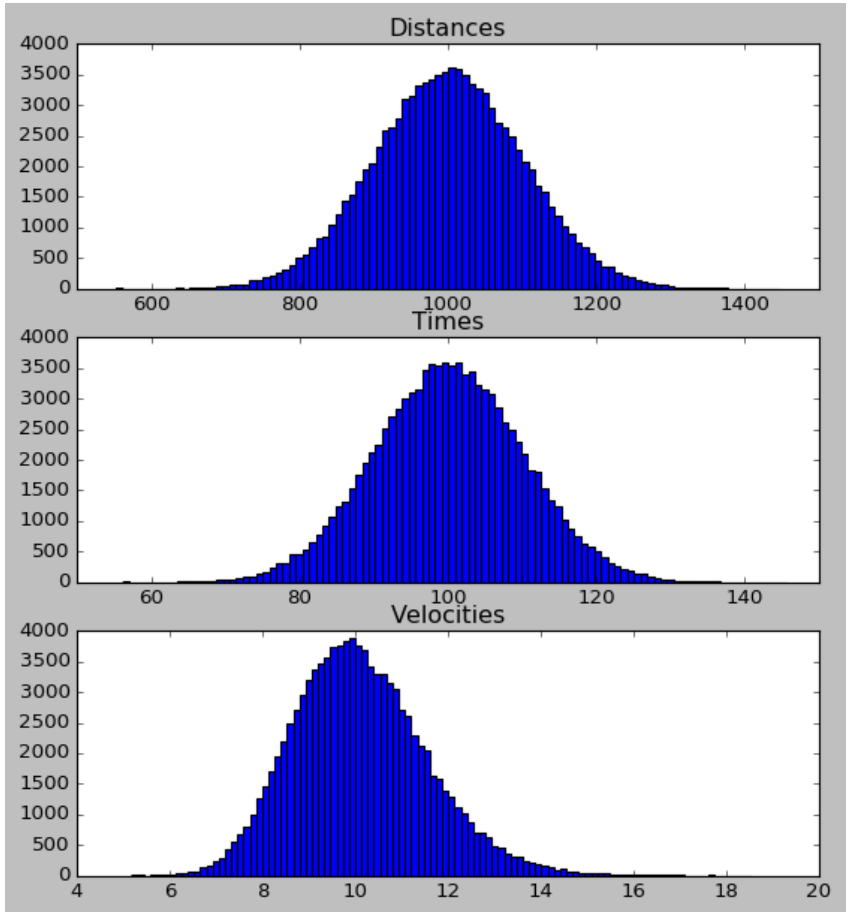
- repeat data evaluation many times with input data sampled from distribution estimating uncertainty

3. Correctly Interpret Your Data

Tip 7: Keep Numbers Behind Error Bars

Example of Uncertainty Quantification: (low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 10 %
- velocity uncertainty $\sqrt{10^2 + 10^2} = 14.36\,\%$, assuming no correlation and normal distribution
- distance error dominates



Monte-Carlo Simulations

- repeat data evaluation many times with input data sampled from distribution estimating uncertainty

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

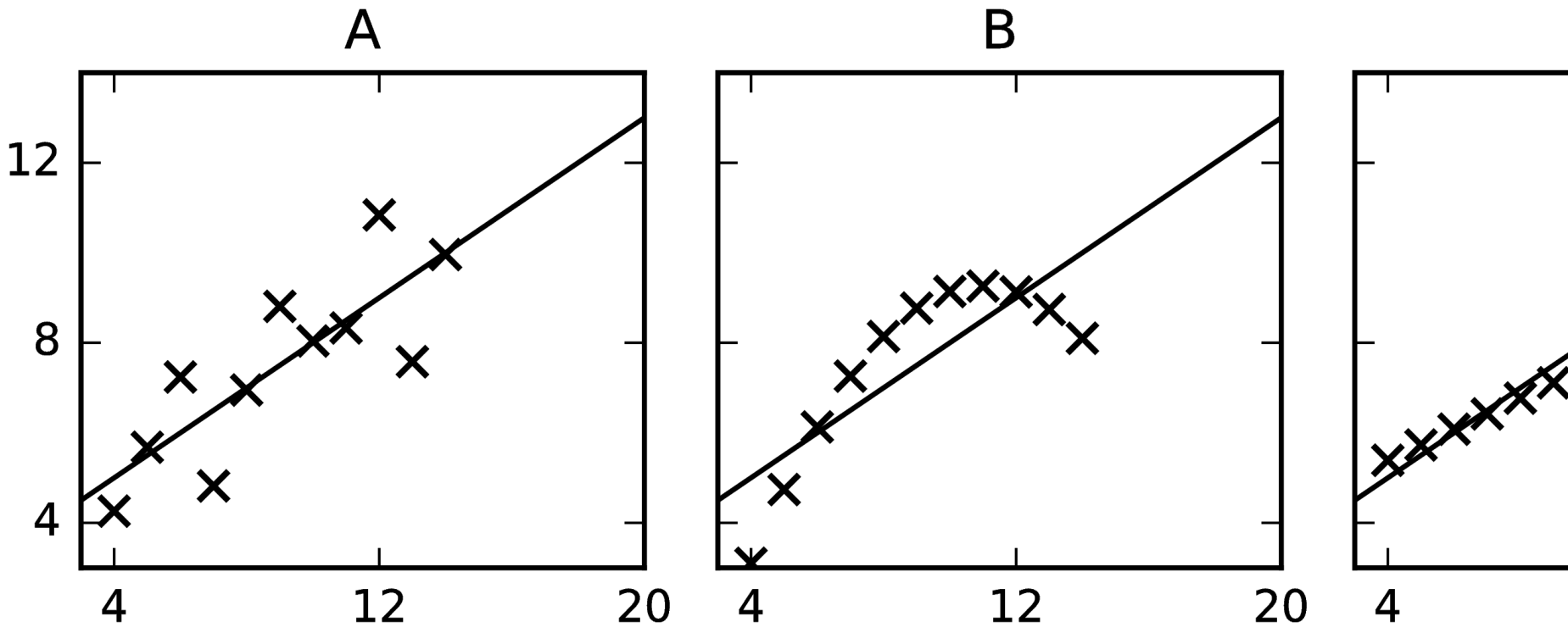
(x, y) data series A to D:

	Series A	Series B	Series C	Series D
$\bar{x}$	9.0	9.0	9.0	9.0
s_x^2	11.0	11.0	11.0	11.0
$\bar{y}$	7.5	7.5	7.5	7.5
s_y^2	4.127	4.128	4.122	4.123
$\rho_{x,y}$	0.816	0.816	0.816	0.817

- These (x, y) data series are pretty much the same. Or are they?

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone



- Anscombe’s quartet [Anscombe, 1973]

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

- Training/Calibration: This is one flower.
  Black rod (flashlight) is a length reference

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

- Training/Calibration: This also is one flower.

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

Continue training and validation, maybe cross-validation …

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

- Analysis result: 4 dandelion flowers. ??

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Let’s count dandelions. Automatically, of course.

- Analysis result: 5 dandelion flowers. ???

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Lessons Learned so far

- don’t blindly trust in numbers
- assess the actual data
- summarizing data is simplification and loses information

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Lessons Learned so far

- don’t blindly trust in numbers
- assess the actual data
- summarizing data is simplification and loses information

Tips:

- quick & dirty plots OK for data assessment
- choose the right way to look at it (scatter plots, time series, spatial data)

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don’t calculate just because something is a number

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don’t calculate just because something is a number
  - temperature –20°C is not “twice as cold as” –10°C

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don’t calculate just because something is a number
  - temperature –20°C is not “twice as cold as” –10°C
  - correlation coefficient 0.9 is not “50 % better than” 0.6

3. Correctly Interpret Your Data

Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don’t calculate just because something is a number
  - temperature –20°C is not “twice as cold as” –10°C
  - correlation coefficient 0.9 is not “50 % better than” 0.6
  - prostate carcinoma with Gleason score 4 is not “twice as large/dangerous/… as” score 2

3. Correctly Interpret Your Data

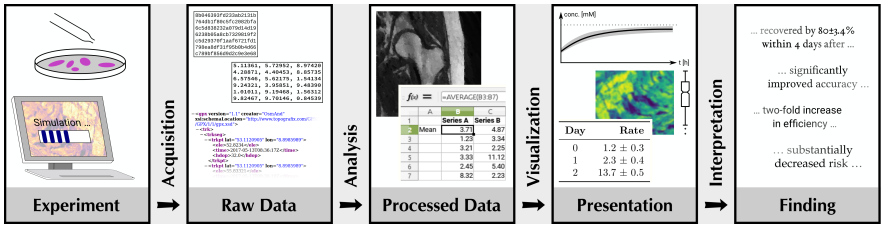
Tip 8: Don’t Rely on Numbers Alone

Don’t Over-Interpret Numbers

- One number means one thing, don’t put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don’t calculate just because something is a number
  - temperature –20°C is not “twice as cold as” –10°C
  - correlation coefficient 0.9 is not “50 % better than” 0.6
  - prostate carcinoma with Gleason score 4 is not “twice as large/dangerous/… as” score 2
- Distinguish extensive and intensive quantities
  - 0.5 l water at 30°C and 0.5 l at 50°C is 1 l at 40°C

4. Present Your Data in a Useful Way

Contents



4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)
- wheat cultivation with 20 % higher yields (relative difference)

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)
- wheat cultivation with 20 % higher yields (relative difference)

Q: What is the baseline?

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

How to Present Data?

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

How to Present Data?

- plot—but that is not the only way
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

How to Present Data?

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines
- tables

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

How to Present Data?

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines
- tables
- number in text

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

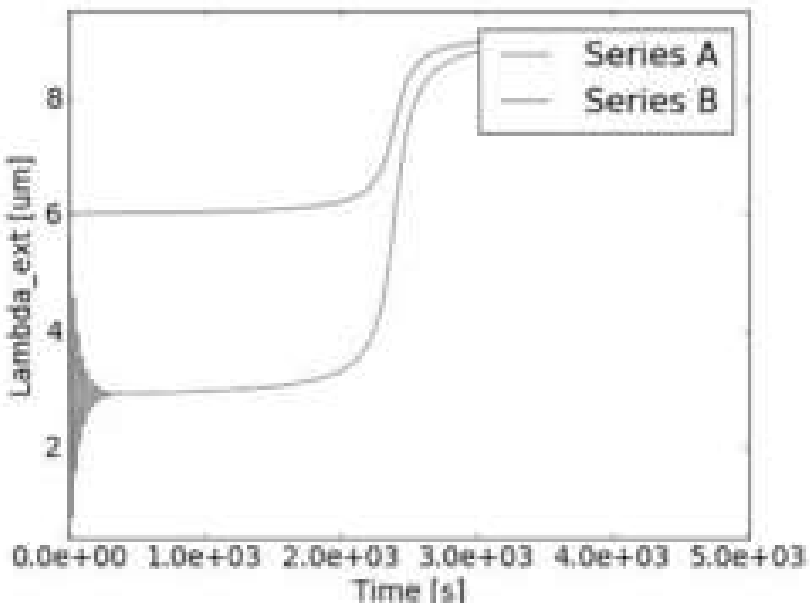
How to Present Data?

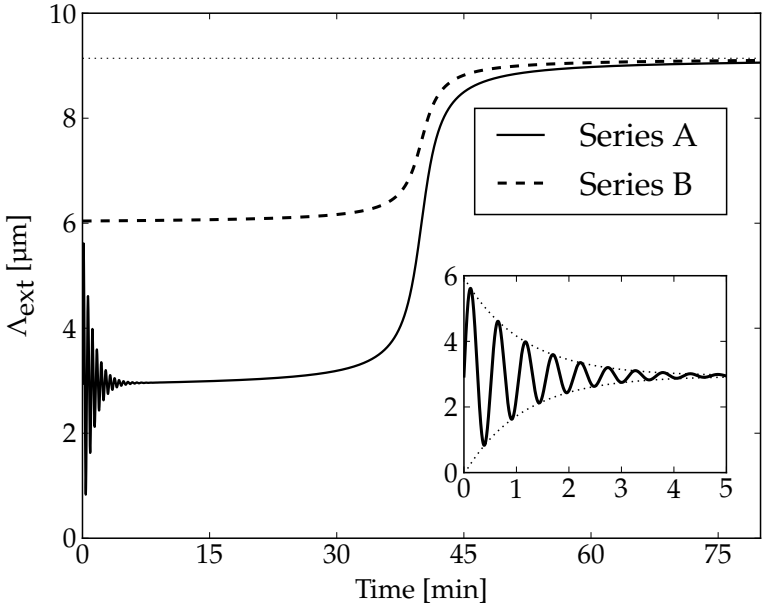
- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines
- tables
- number in text
- spatial or other visualization

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Ugly vs. Nice Plot: What is wrong on the left?

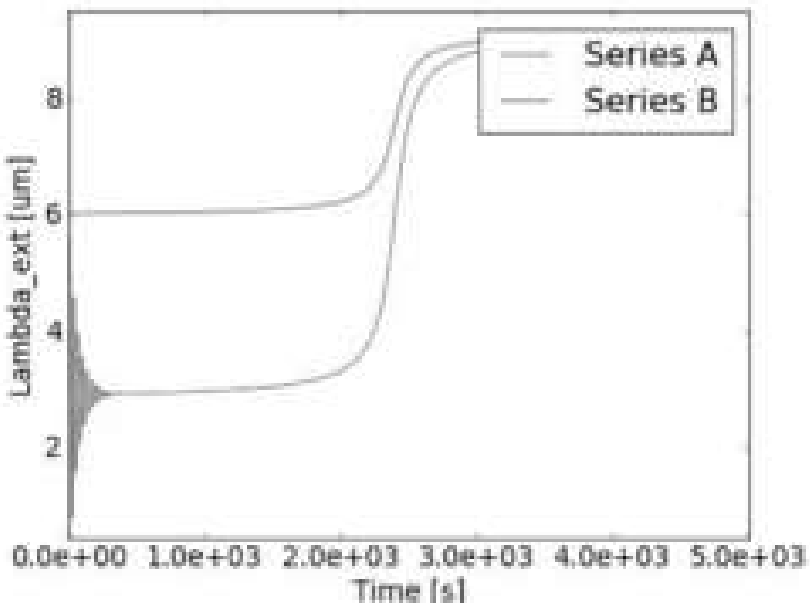


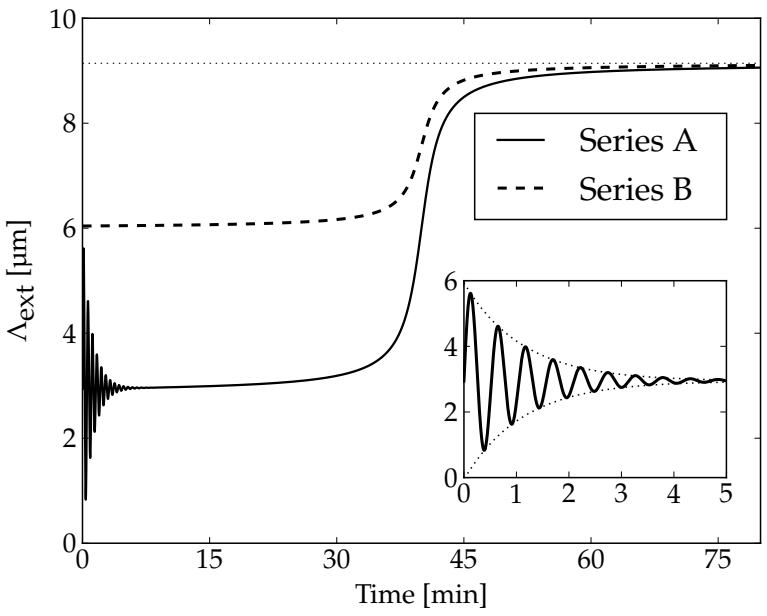


4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Ugly vs. Nice Plot: What is wrong on the left?



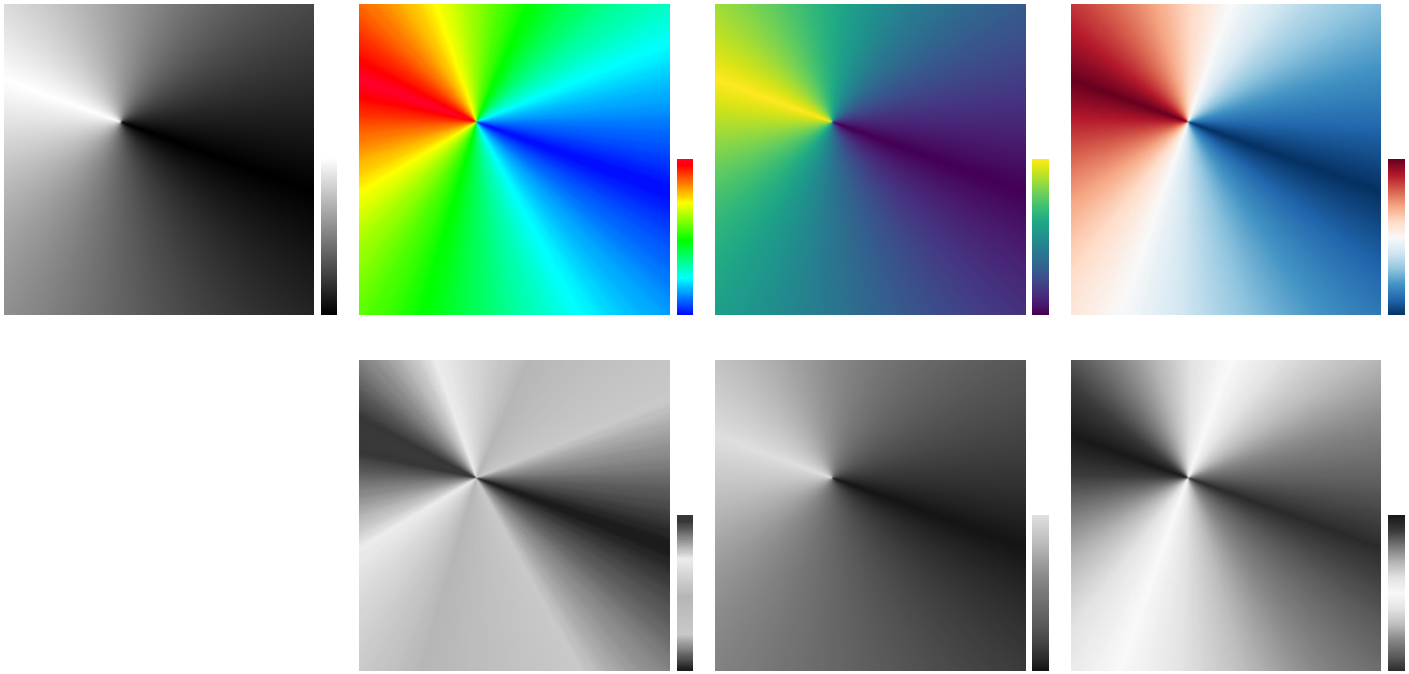


text ugly, numbers hard to read and not intuitive, curves not distinguishable, legend occludes data, oscillation not visible, value range tries but fails to start at zero, image is blurred and low-quality jpeg, relevant information missing

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Color Scales



- distinguish sequential and divergent (and categorical) color scales
- consider color-blind people and crappy b/w printers
- avoid rainbow colormap

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like “open data”
- ensures long-term availability (good scientific practice)
- some journals require sharing data

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like “open data”
- ensures long-term availability (good scientific practice)
- some journals require sharing data
- preparing (cleaning up, standardizing, …) and documenting data may reveal mistakes

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like “open data”
- ensures long-term availability (good scientific practice)
- some journals require sharing data
- preparing (cleaning up, standardizing, …) and documenting data may reveal mistakes
- people might use the data
  - yourself at different lab
  - collaborations with others might evolve
  - others might find gems you will never go looking for

4. Present Your Data in a Useful Way

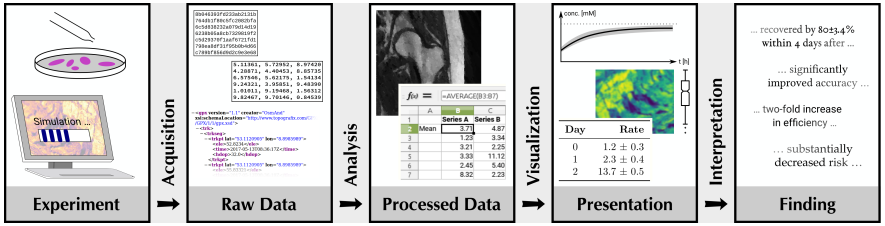
Tip 10: Share Your Data

How?

- own, institutional, project website (not optimal)
- domain-specialized or general purpose data repositories
- supporting information

5. Summary

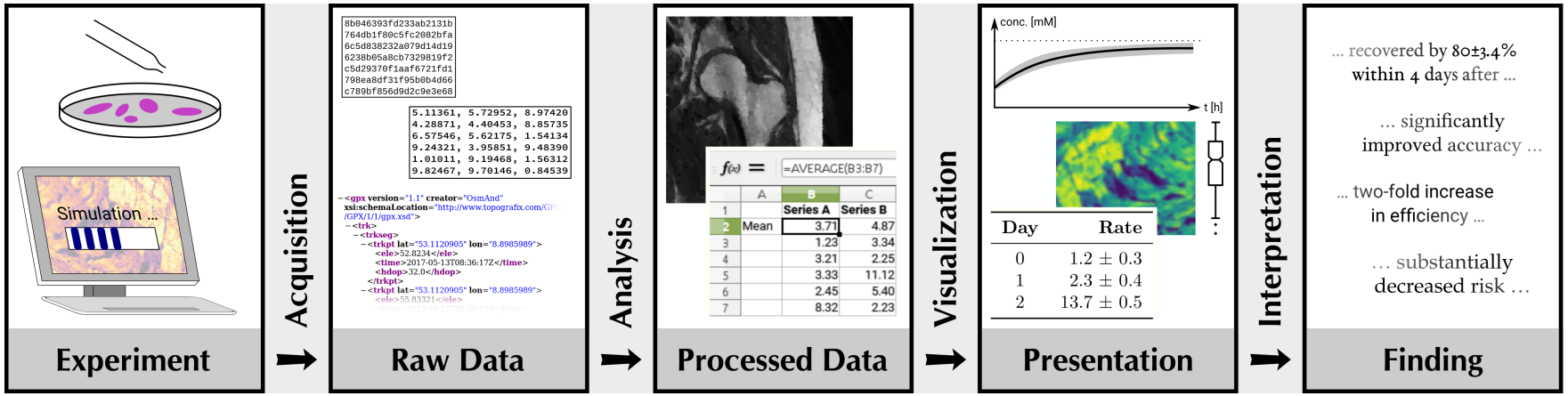
Contents



5. Summary

Tips & Tricks for Getting Most Value out of Numerical Data

1. Plausibility-Check Your Data
2. Track Your Sources
3. Beware of Computational Pitfalls
4. Treat Units With Respect
5. Verify Your Formulas
6. Know Thy Statistical Methods
7. Keep Numbers Behind Error Bars
8. Don’t Rely on Numbers Alone
9. Properly Present Data
10. Share Your Data



Contact: Ole Schwen ‹ole.schwen@mevis.fraunhofer.de›